PATENTTI- JA REKISTERIHALLITUS
NATIONAL BOARD OF PATENTS AND REGISTRATION

PCT/FI 03/00195

10/507144

Helsinki 16.5.2003



# ETUOIKEUSTODISTUS
# PRIORITY DOCUMENT

| | |
|---|---|
| Hakija<br>Applicant | **Master´s Innovations Ltd Oy**<br>**Helsinki** |
| Patenttihakemus nro<br>Patent application no | **20020532** |
| Tekemispäivä<br>Filing date | **20.03.2002** |
| Kansainvälinen luokka<br>International class | **G06F** |
| Keksinnön nimitys<br>Title of invention | **"Menetelmä ja laitteisto datan kääntämiseksi"** |

Täten todistetaan, että oheiset asiakirjat ovat tarkkoja jäljennöksiä Patentti- ja rekisterihallitukselle alkuaan annetuista selityksestä, patenttivaatimuksista, tiivistelmästä ja piirustuksista.

This is to certify that the annexed documents are true copies of the description, claims, abstract and drawings originally filed with the Finnish Patent Office.

**PRIORITY DOCUMENT**
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

Marketta Tehikoski
Apulaistarkastaja

Maksu 50 €
Fee 50 EUR

*Maksu perustuu kauppa- ja teollisuusministeriön antamaan asetukseen 1027/2001 Patentti- ja rekisterihallituksen maksullisista suoritteista muutoksineen.*

*The fee is based on the Decree with amendments of the Ministry of Trade and Industry No. 1027/2001 concerning the chargeable services of the National Board of Patents and Registration of Finland.*

Osoite: Arkadiankatu 6 A, P.O.Box 1160, FIN-00101 Helsinki, FINLAND
Puhelin: 09 6939 500
Telephone: + 358 9 6939 500
Telefax: 09 6939 5328
Telefax: + 358 9 6939 5328

L1

**Menetelmä ja laitteisto datan kääntämiseksi – Metod och apparatur för att transformera data**

Keksintö koskee yleisesti datan luokittelua ja kääntämistä tai muuntamista toiseen alkuperäistä vastaavaan muotoon. Erityisesti keksintö koskee kielen kääntämistä.

Luonnollisten kielten automaattiseen kääntämiseen käytetään nykyisin pääasiassa kahta tekniikkaa: konekäännös- ja käännösmuistitekniikkaa. Käännettävää kokonaisuutta kutsutaan yleisesti syötetietovirraksi ja syötetietovirta sisältää tunnistettavissa olevia elementtejä. Luonnollisen kielen tapauksessa syötetietovirta sisältää siis lauseita ja/tai virkkeitä ja tunnistettavat elementit ovat sanoja mahdollisine etu- ja jälkiliitteineen.

Konekäännöstekniikassa syötetietovirran elementit analysoidaan hyvin tarkasti määritetyn säännöstön mukaisesti. Analysoiduista elementeistä tuotetaan järjestelmään ohjelmoitujen, tuhansien jäsennyssääntöjen avulla alkuperäistä lausetta tai virkettä vastaava jäsennyspuu, joka kuvaa elementtien riippuvuutta toisistaan ja toisista alipuista. Esimerkiksi lauseen "kissa kävelee" elementti "kissa" tulkitaan subjektiksi, joka riippuu predikaatista "kävelee". Nämä riippuvuussuhteet määritetään yksinkertaistettujen sääntöjen mukaan edeten yleisistä yksityiskohtaisempiin, esimerkiksi tässä esimerkkilauseessa aluksi tarkastellaan kokonaista virkettä, joka koostuu tässä yhdestä lauseesta. Lause sisältää predikaatin ja niin sanotun nominaalifraasin. Tämä nominaalifraasi sisältää subjektin ja mahdolliset sitä kuvaavat adverbiaalit. Lauseen subjekti on substantiivin nominatiivi ja yksikkö, predikaatti on verbin preesens ja yksikkö. Näin tuotettu jäsennyspuu muunnetaan sitten kohdekielen jäsennyspuurakenteeksi erillisten muunnossääntöjen avulla. Kohdekielen jäsennyspuurakenteesta tuotetaan eri vaiheiden jälkeen kohdekielisen lauseen tai virkkeen rakenteen mukainen elementeistä koostuva kokonaisuus. Käännöksen tuottamiseksi on siis käytettävä vähintään kolmea eri sääntökantaa jäsennyspuiden tuottamiseen, muuntamiseen ja generoimiseen, sekä joukkoa erillisiä analysointi- ja generointisääntökantoja tai muita vastaavia mekanismeja.

Käännösmuistitekniikassa elementtejä ei analysoida, vaan syötetietovirran kokonaisia lauseita tai virkkeitä verrataan tietokannassa oleviin elementtijonoihin merkkijonovertailuna. Jos samanlainen merkki- tai elementtijono löydetään, sen käännös on tähän jonoon assosioitu vastinkielinen merkki- tai elementtijono, ja se tulostetaan vasteena syötetietovirran käännöspyyntöön. Käännösmuistitekniikkaa hyödyn-

tävät järjestelmät ovat tehokkaimmillaan, kun saman tekstin eri versioita käännetään uudestaan tai kun käännettävät tekstit sisältävät samoja lauseita. Olemassa olevista tekniikoista käännösmuisti on melko tehokas ja käyttökelpoinen poistamaan rutiinityötä. Käännösmuistit eivät kuitenkaan kykene kääntämään tarpeeksi tarkasti aiemmasta poikkeavia lauseita, vaan kääntäjä joutuu muokkaamaan tekstiä aina, kun se sisältää uuden kääntämättömän lauseen.

Konekäännöstekniikkaa voidaan soveltaa niin sanotussa esimerkkiperusteisessa konekäännöksessä (example-based machine translation, EBMT), jonka perusidea on se, että käännetään syötevirke matkimalla samantapaisten valmiiden esimerkkien käännöksiä. Esimerkkiperusteisessa konekäännöksessä yritetään siis tuottaa lopputulos yhdistämällä kahden eri käännöksen osia yhdistämällä niiden jäsennyspuita syötetietovirtaa vastaavaksi jäsennyspuuksi. Muita tunnettuja tapoja perinteisen konekäännöstekniikan ongelmien kiertämiseksi ovat muistiperusteinen (memory-based MT), analogiaperusteinen (analogy-based MT) ja tapausperusteinen (case-based MT) konekääntäminen.

Tilastolliset käännösjärjestelmät perustuvat sanojen esiintymisen todennäköisyyteen valmiissa käännöksissä. Esimerkiksi voidaan etsiä vastaavuudet alkuperäiskielisistä ja käännetyistä virkkeistä, ja laskea todennäköisyys sille, kääntyykö alkuperäinen sana yhdeksi vai kahdeksi sanaksi vai jääkö se käännöksestä kokonaan pois. Tämän perusteella tuotetaan käännössäännöt.

On myös olemassa erinäisiä rajoitettuihin kieliin tai alikieliin perustuvia järjestelmiä. Niiden käyttö on kuitenkin hyvin kurinalaista, sillä käyttäjän antaman syötteen on oltava tarkoin määriteltyjen sääntöjen mukaista. Tämä vaatii erityistä mukautumiskykyä ja –halua käyttäjältä. Koulutettu käyttäjä pääsee kuitenkin lähelle ideaalista tulosta tällaisessa rajoitetussa järjestelmässä, eikä käyttäjän apua yleensä käännösvaiheessa tarvita.

Tunnetun tekniikan mukainen konekääntäminen edellyttää monimutkaisten säännöstöjen ja semantiikan ohjelmointia, jotta yksittäisten sanojen lauseyhteydet saadaan esille. Tämä vaatii edelleen raskasta ohjelmointia ja tyypillisesti vielä ammattilaisen tulkintaa. Esimerkki-, muisti-, analogia- tai tapausperusteisten konekäännösten soveltaminen vaatii useiden vaikeasti toteutettavien osavaiheiden suorittamista. Tarvitaan alkuperäisen ja käännöskielisen kielen jäsennyspuut, jotta voidaan etsiä ja ohjelmoida virkkeiden vastinosapuut. Tämä asettaa vaatimuksensa tiedon esitysmuodolle ja tuotetut puurakenteet ovat aina raskaita toteuttaa ja käyttää.

Jos käännösmuistijärjestelmä ei voi tuottaa käännöstä käyttäjän syötteeseen, se joko antaa vaihtoehtoisia tuloksia, joista käyttäjä voi valita haluamansa tai pyytää käyttäjää syöttämään oikean käännöksen. Usein käyttäjä muuttaa käännösvirkkeen rakennetta niin paljon, että käännösmuistijärjestelmään tallennetaan vain kokonaisen virkkeen tai lauseen käännösvaste. Käännösjärjestelmien opettamiseen tarvitaan tyypillisesti suuri määrä oikeanlaisia valmiita käännöksiä. Käännösmuistitekniikan ongelmana on sen kyvyttömyys kääntää aivan uusia, aiemmin kääntämättömiä lauseita. Ongelmaa on yritetty ratkaista yhdistämällä tunnettuja käännöksiä uusiin syötteisiin, muun muassa neuraaliverkkoja ja tilastollisia todennäköisyyksiä hyväksi käyttäen. Tulokset eivät kuitenkaan ole olleet lupaavia, sillä käännösmuistit eivät kykene muokkaamaan tarkasti oikeaa tulosta samankaltaisen lauseen perusteella, vaan yleensä kopioivat syötelauseelle lähimmän vastaavan käännösvasteen sellaisenaan lopputulokseksi.

Kaupallisesti käännösmuistitekniikkaa käyttävät tuotteet ovat menestyneet konekäännöstekniikkaa hyödyntäviä paremmin, koska jälkimmäinen vaatii raskasta prosessointia ja siten laitteet ovat tyypillisesti joko liian hitaita tai liian kalliita. Molempien tekniikoiden kaupallistamisen ongelmana on suuri työmäärä sovitettaessa järjestelmiä uusille toimialoille tai mukautettaessa niitä kielen rakenteiden ja sanaston kehittyessä.

Keskeiset ongelmat olemassa olevien ratkaisujen takana ovat koneilta vaadittava tehokkuus ja nopeus sekä menetelmän kattavuus eli se, kuinka suuri osa käännöksistä on riittävän hyviä. Nämä kaksi ovat lisäksi sidoksissa toisiinsa. Periaatteessa käännösjärjestelmän pitäisi kyetä kääntämään miljardeja mahdollisia lauseita, jotka syntyvät kymmenien tuhansien sanojen lukuisista erilaisista kombinaatioista. Esimerkkipohjaisissa järjestelmissä tätä valtavaa vaihtoehtojen määrää pyritään hallitsemaan tallentamalla paljon esimerkkejä, joista jokaista voidaan sovittaa moneen käännettävään tekstiin. Esimerkiksi 10 000 esimerkkiä, joista jokainen sopii 10 000 käännettävään kohteeseen, kykenee käsittelemään $10\ 000^2 = 0{,}1$ miljardia potentiaalista käännettävää lausetta. Lisäksi esimerkkipohjaisissa järjestelmissä voidaan soveltaa segmentointia, eli jakaa käännettävä syöte pienempiin osiin, jolloin erilaisia kombinaatioita on vähemmän. Tältä pohjalta esimerkkipohjaisten käännösjärjestelmien ongelmakokonaisuus voidaan ryhmitellä esimerkiksi seuraavaan neljään osaongelmaan:

1. Esimerkkien määrä. Käännösjärjestelmän täytyy kyetä hallitsemaan suurta määrää esimerkkejä tehokkaasti, sekä kyetä hakemaan sopivia esimerkkejä nopeasti suurista tietokannoista. Tähän pystyvät perinteiset käännösmuistit, mutta eivät

jäsennyspuita tai muita tekstimuotoa monimutkaisempia esitysmuotoja käyttävät konekäännösjärjestelmät tai vastaavia tekniikoita käyttävät esimerkkipohjaiset käännösjärjestelmät.

2. Esimerkkien yleistys, haku ja sovitus. Yhden esimerkin tulee sopia moneen käännettävään kohteeseen (lähdekielen lauseeseen tai sen osaan), sopivan esimerkin haun tietokannasta on oltava nopea ja sovituksen tehokas. Käännösmuistit eivät tähän kykene, sillä ne sovittavat kohteen vain tekstivertailulla eivätkä kykene yleistykseen. Sen sijaan monet esimerkkipohjaiset järjestelmät pystyvät sovittamaan saman esimerkin moneen käännettävään kohteeseen soveltamalla kieliteknologiaa. Niissä sovitus on yleensä monivaiheinen, käyttää laskennallisesti hankalia menetelmiä, hitaita ja monimutkaisia hakuja sekä rajaavia heuristiikkoja, jolloin niiden skaalattavuus on huono, eli osaongelma 1 ei ratkea.

3. Segmentointi ja segmenttien yhdistely. Jos teksti käännetään sana kerrallaan, tarvittavien esimerkkien määrä on pieni, mutta käännöksen laatu erittäin huono. Jos esimerkin (segmentin) koko on lause tai virke, käännös voidaan yleensä tehdä laadukkaasti, mutta tarvittavien esimerkkien määrä nousee miljardeihin (ilman sovitusta – kts. osaongelma 2). Tarvittavien esimerkkien määrää voidaan pienentää oleellisesti käyttämällä lausetta lyhyempiä segmenttejä. Tällöin segmenttien yhdistely tulee uudeksi ongelmaksi ja epätarkkojen käännösten osuus lisääntyy. Aina kokonaisen esimerkkilauseen tai virkkeenkään käyttö ei takaa oikeellisuutta, sillä lauseen/virkeen oikea tulkinta voi vaatia jopa lauseyhteyden tai kappalleen ulkopuolista kontekstia tai semanttista maailmanmallia. Erityistä tulkintaa vaaditaan esimerkiksi runoja käännettäessä. Riippuen käytettävästä yleistystekniikasta (osaongelma 2) "turvallisen" segmentoinnin tekeminen voi olla helpompaa. Toisaalta usein riski väärästä käännöksestä lisääntyy.

4. Käännösvasteen muokkaaminen. Jos esimerkkipohjaisessa käännösjärjestelmässä käytetään vain käännösesimerkkejä ja niiden käännösvasteita tekstimuotoisina, ilman segmentointia, ei lähdekielisen tekstin käännösvastetta tarvitse muokata. Jos käytetään "turvallista" segmentointia (osaongelma 3), käännösvaste voidaan tehdä yhdistämällä segmenttien käännökset. Jos taas käytetään yleistystä (osaongelma 2), tai lyhyiden segmenttien yhdistelyä, käännösvasteen muokkaaminen voi olla hyvin hankalaa.

Tunnetuilla menetelmillä kaikkien näiden neljän osaongelman ratkaisu ei ole onnistunut samalla kertaa eli kokonaisuus ei toimi. Käännösmuistijärjestelmät ratkaisevat osaongelmat 1 ja 4, mutta keinojen puuttuessa osaongelmaan 2 niiltä puuttuu yleis-

tettävyys. Tutkimuksellisissa esimerkkipohjaisissa käännösjärjestelmissä esitetään ratkaisumalleja osaongelmaan 2. Esimerkiksi tunnettu käännösohjelma ReVerb (Collins, B., Cunningham, P., Veale, T., An Example-Based Approach to Machine Translation, Proc. of AMTA conference, October 1996, pp.1-13) pyrkii ratkaisemaan osaongelmat 2 ja 4 yleistämällä esimerkkejä sanojen lauseenjäsennyksen avulla ja ottamalla käytettävän esimerkin valinnassa huomioon käännösvasteen muokattavuuden. Sen käyttämän haku- ja sovitusmekanismin monimutkaisuus ja parinsadan esimerkin tietämyskanta eivät kuitenkaan näytä skaalautuvan osaongelman 1 ratkaisemiseksi. Pangloss (Brown, R.D., Example-Based Machine Translation in the Pangloss System, Proceedings of the 16th International Conference on Computational Linguistics, August 1996) taas käyttää hybridimallia, jossa pohjana on tekstipohjaisen käännösmuistin ratkaisu osaongelmaan 1, jonka yleisyyttä on lisätty käyttämällä esimerkiksi päivämäärien kääntämiseen sovituspohjia, jotka tunnistavat ja kääntävät kaikki päivämäärät. Tämä malli on suhteellisen turvallinen osaongelman 4 suhteen, mutta sen yleistettävyys (osaongelma 2) jää suhteellisen vähäiseksi, sillä kaikkia syötteitä ei kyetä kääntämään. Pangloss käyttääkin erillistä konekäännösjärjestelmää kääntääkseen loput syötteet ja saavuttaakseen riittävän yleistettävyyden. Kaupallisesti parhaiten menestynyt tuote, Trados (http://www.trados-com), ratkaisee käännösmuistina osaongelman 1 ja yrittää soveltaa neuraalilaskentaa osaongelman 2 ratkaisemiseen. Tässä ei kuitenkaan onnistuta, sillä neuraalilaskenta ei riitä osaongelmaan 2 ja, ennen kaikkea, osaongelma 4 jää ratkaisematta, samoin 3. Yleensäkkään näissä järjestelmissä ei juuri kyetä hyödyntämään segmentointia, poikkeuksena lähinnä Pangloss, jossa keskimääräinen segmentti on noin kolmen sanan pituinen niille syötteille, joita se kykenee käsittelemään.

Keksinnön tavoitteena on tuottaa tehokas, joustava menetelmä ja järjestely datan luokittelemiseksi ja edelleen kääntämiseksi. Lisäksi keksinnön tavoitteena on tuottaa käännösjärjestely, joka on helposti mukautettavissa uudenlaisiin syötetietovirtoihin ja rakenteisiin.

Tavoite saavutetaan siten, että dataa käsitellään sopivan kokoisina segmentteinä, tehokkailla analysointimenetelmillä. Jokainen segmentti saa analysointitulosten perusteella yksikäsitteisen luokituksen, jota voidaan käyttää erittäin tehokkaasti segmenttien vertailuun ja suurten tietämyskantojen hakuavaimena. Tehokkuuden ansiosta tietämyskannan kokoa ja esimerkkien määrää voidaan lisätä edelleen, mikä parantaa kattavuutta ja laatua.

Keksinnölle on tunnusomaista se, mitä sanotaan itsenäisten patenttivaatimusten tunnusmerkkiosissa. Keksinnön edullisia suoritusmuotoja on kuvattu epäitsenäisissä patenttivaatimuksissa.

Keksinnön edullisen suoritusmuodon mukaan syötetietovirran kääntäminen toiseen muotoon tapahtuu vaiheittain. Keksinnön edullisen suoritusmuodon mukaisessa menetelmässä käytetään hyväksi sinänsä tunnettuja menetelmiä syötetietovirran segmentoimiseksi eli jakamiseksi osiin. Käyttökelpoisia segmentointimenetelmiä ovat esimerkiksi syötetietovirran segmentointi välimerkkien avulla, lauseina, fraaseina tai välikesanojen avulla, vaikkapa katkaisemalla segmentti ja-sanan jälkeiseen sanaan tai ennen sivulauseen aloittavia sanoja. Keksinnön erään edullisen suoritusmuodon mukaan käytetään sellaista segmentointimenetelmää, jossa syötteen jako segmentteihin tehdään siten, että muodostetut segmentit löytyvät mahdollisimman kattavasti jo tietämyskannassa olevista segmenteistä.

Keksinnön edullisen suoritusmuodon mukaan aluksi yritetään kääntää syötetietovirtaa mahdollisimman vähän resursseja kuluttavasti, esimerkiksi käännösmuistitekniikan avulla. Tyypillisesti ainakin osa syötetietovirrasta saadaan käännettyä suoraan ja nopeasti. Syötetietovirran jäljelle jääneelle osalle tehdään kevyt analysointi, jossa syötetietovirran elementeille tuotetaan kullekin jokin analyysitulos. Tässä hakemuksessa yksittäisen elementin kohdalla puhutaan analyysituloksesta, koko segmenttiä koskevaa analyysitulosta sanotaan luokitteluksi. Luokitus muodostetaan analyysituloksista, esimerkiksi katenoimalla, eli liittämällä yhteen, elementtien analyysitulokset ja niiden väliin lisätyt välikesymbolit yhtenäiseksi merkkijonoksi. Tätä segmentin luokitusta verrataan tietämyskannassa olevien segmenttien luokituksiin tehokkaan indeksi- tai tietokantahaun avulla. Haun tuloksena tietämyskannasta palautetaan segmentit, joilla on sama tai lähes sama luokitus kuin syötetietovirran segmentillä. Näistä tietämyskannan segmenteistä valitaan yksi syötetietovirran segmenttiä parhaiten vastaava segmentti tiettyjen sääntöjen perusteella. Segmenteistä voidaan valita esimerkiksi se, jossa on eniten samoja elementtejä kuin käännettävässä syötetietovirran osassa.

Käännöksen tuloksena palautetaan tietämyskannasta parhaiten syötetietovirran segmenttiä vastaavaan segmenttiin assosioitu vastinsegmentti. Syötetietovirran segmentin sanat, joita ei ollut tässä parhaiten vastaavassa segmentissä, käännetään erikseen jollain tunnetulla tekniikalla, esimerkiksi generoimalla sana kerrallaan sopiva taivutusmuoto sanakirjasta löydetylle vastinelementille. Keksinnön mukainen luokittelu ja segmenttien vertailu tietämyskannan segmentteihin tuottaa hyviä tuloksia tehokkaasti jo melko pienestäkin tietämyskannasta.

Keksinnön mukainen menetelmä poikkeaa huomattavasti tunnetusta konekäännöstekniikasta, koska keksinnössä ei esimerkiksi muodosteta jonkin kieliopin tai säännöstön mukaista jäsennyspuuta syötetietovirrasta. Myöskään sääntöjä ei keksinnön mukaiseen menetelmään tarvitse ohjelmoida. Lisäksi keksinnön mukaisesti syötetietovirran elementtejä verrataan tietämyskannan elementteihin myös sellaisenaan, kun tunnetuissa konekäännöstekniikoissa elementtejä käsitellään aina analysoituina.

Keksinnön mukainen menetelmä poikkeaa käännösmuistitekniikoista ja esimerkkipohjaisista käännösjärjestelmistä tarjoamalla ratkaisun kaikkiin neljään esimerkkipohjaisten käännösjärjestelmien ongelmakokonaisuuteen. Käännettävän syötesegmentin analyysituloksesta muodostettu luokitus toimii hakuavaimena, jolla haetaan tietämyskannasta siihen sovellettavan esimerkkikäännöksen lähdekielen segmentti (ratkaisee osaongelmat 1 ja 2). Haku on erittäin tehokasta, sillä siihen voidaan soveltaa indeksointi- ja tietokantatekniikoita monimutkaisten puuvertailujen ja aktivointijärjestelyjen sijaan. Linkitys esimerkkikäännöksen kohdekielen segmenttiin muokkaa käännösvastetta varsin turvallisella menetelmällä (ratkaisee paljolti osaongelman 4). Osaongelmien 1 ja 2 ratkettua nykyisin tunnettuja menetelmiä paremmin tietämyskannan kokoa voidaan kasvattaa suureksi tehokkuuden kärsimättä oleellisesti, mikä parantaa edelleen kattavuutta. Siksi tietämyskantaan voidaan myös lisätä lyhyitä ja pitkiä segmenttejä samoistakin esimerkeistä. Käännösten laatu taataan käyttämällä mahdollisimman pitkiä segmenttejä, jotka ovat turvallisempia (3 ja 4) samalla kun lyhyet segmentit takaavat yleistettävyyden ja kattavuuden paremmin kuin esimerkiksi neuraalimenetelmä tai sanakirjasovitus. Näin segmentointia voidaan hyödyntää käyttämällä tilanteeseen sopivaa segmenttikoa (osaongelma 3).

Tekstimuotoisten luonnollisten kielien ja formaalien kielien kääntämisen lisäksi keksinnön edullisia suoritusmuotoja voidaan käyttää useilla tiedon luokittelua ja muuntamista soveltavilla alueilla. Tekstimuotoisen syötetietovirran käsittelyn lisäksi keksinnön erästä edullista suoritusmuotoa voidaan käyttää myös puhetta tulkattaessa. Kun käännös tehdään ohjelmointikielestä toiseen, on kääntäminen luonnollisesti paljon kurinalaisempaa ja syntaksien mukaista.

Keksinnön mukainen menetelmä on nykyisiä suorituskykyisempi, koska sen vasteaika on oleellisesti nykyratkaisuja parempi. Lisäksi keksinnön mukaiset menetelmät ovat hyvin mukautuvia eli niitä käyttämällä saadaan oikeita tulosvirtoja aiempaa suuremmassa osassa tapauksista oleellisesti aiempaa nopeammin. Tehokkuuden ansiosta myös tietämyskannan kokoa ja esimerkkien määrää voidaan kasvattaa, mikä parantaa edelleen kattavuutta. Tehokkuuden takia menetelmän ei myöskään tarvitse käyttää lisäheuristiikkoja tai rajoituksia, jotka voivat itse asiassa huonontaa suori-

tuskykyä, esimerkkinä rajautuminen segmentoinnissa jäsennyspuun alipuihin tai predikaattien poikkeava käsittely hakurakenteissa. Menetelmä ei kuitenkaan estä tällaisten heuristiikkojen tai lisäysten käyttöä silloin, kun ne ovat hyödyllisiä. Menetelmä on kääntämisen lisäksi helposti yleistettävissä muidenkin sovellusten käyttöön, kuten ohjelmointikielikonversioihin ja monikanavajulkaisuihin.

Seuraavassa keksintöä ja sen edullisia suoritusmuotoja selostetaan tarkemmin oheisten kuvioiden avulla, joissa

kuvio 1 esittää lohkokaaviona keksinnön erään edullisen suoritusmuodon mukaista laitteistojärjestelyä,

kuvio 2 esittää keksinnön erään edullisen suoritusmuodon mukaista käsiteltävää syötetietovirran osaa,

kuvio 3 esittää keksinnön erään edullisen suoritusmuodon mukaisen tietämyskannan osan rakennetta,

kuvio 4 esittää keksinnön erään edullisen suoritusmuodon mukaista tulostietovirran osaa,

kuvio 5 esittää vuokaaviona keksinnön erään edullisen suoritusmuodon mukaista menetelmää datan luokittelemiseksi,

kuvio 6 esittää vuokaaviona keksinnön erään edullisen suoritusmuodon mukaisen tietämyskannan kasvattamista, ja

kuvio 7 esittää vuokaaviona keksinnön erään edullisen suoritusmuodon mukaista datan kääntämistä.

Kuviossa 1 on esitetty keksinnön erään edullisen suoritusmuodon mukainen laitteistojärjestely. Näyttö 101 ja näppäimistö 102 toimivat rajapintana käyttäjälle. Massamuistissa 105 säilytetään tietämyskantoja indekseineen, käytettäviä ohjelmia ja sääntöjä. Keskusmuistissa 104 taas säilytetään kulloinkin käsiteltävää osaa syötetietovirrasta ja hakuindeksistä. Lisäksi laitteistossa on prosessori 103, joka käsittelee dataa ja I/O-liityntöjä 106, joiden kautta laitteistoon voidaan liittyä sen ulkopuolelta.

Näytöllä 101 voidaan esittää käyttäjälle suorituksen tuloksia ja/tai vaiheita. Näppäimistön 102 avulla taas käyttäjä voi syöttää laitteistoon varsinaisen syötetietovirran lisäksi vaikkapa vastine-ehdotuksia sanoille ja lauserakenteille, joita järjestelmä

ei osaa kääntää. Kaikki näytöllä 101 esitettävä ja näppäimistöltä 102 syötettävä data käsitellään prosessorissa 103. Prosessoriin 103 liitettyjen I/O-kanavien kautta järjestelmä voi myös olla yhteydessä muihin järjestelmiin ja käyttäjiin sekä lähettää ja vastaanottaa syöte- ja tulostietovirtoja. Keksinnön mukaista järjestelyä voidaan siis käyttää useastakin paikasta ja myös tietoliikenneyhteyden välityksellä.

Keskusmuistissa 104 sijaitsee se osa syötetietovirrasta, jota käsitellään parhaillaan. Lisäksi keskusmuistissa 104 on käsiteltävän syötetietovirran segmentit. Käsiteltävä syötetietovirran osa on ryhmitelty osiin eli segmentteihin tiettyjen sääntöjen perusteella, joita käsitellään myöhemmin tässä hakemuksessa. Järjestelmän massamuistissa 105 on tietämyskanta, jossa ovat segmentit ja niiden vastinsegmentit. Myös elementeille ja niiden vastinelementeille voi olla erillinen tietokanta. Tämä elementtitietokanta voi vastata perinteistä sähköistä sanakirjaa, jossa on sanakohtaiset vastaavuudet tai keksinnön kulloisenkin suoritusmuodon mukaan elementit voivat olla vaikkapa matemaattisia ilmaisuja tai formaalien kielien käskyjä tai parametreja. Massamuistissa 105 on myös erilaisia käsittelysääntöjä, kuten esimerkiksi segmentointisäännöt, joiden perusteella käsiteltävä syötetietovirran osa jaetaan segmentteihin. Lisäksi massamuistissa 105 on muunnossääntöjä esimerkiksi sanajärjestyksen muuttamiseksi segmentin ja sen vastinsegmentin välillä, sekä tarvittavat ohjelmat, kuten esimerkiksi syötetietovirran käsittelemiseksi tarvittavat analysointi- ja generointiohjelmat. Analysointiohjelman avulla syötetietovirran elementeille tuotetaan analyysitulokset. Generointiohjelma puolestaan tuottaa analyysituloksen avulla tulostietovirran elementin. Kuvion 1 laitteistojärjestely on tyypillinen keksinnön mukaiselle järjestelylle, mutta alan ammattilaiselle on ilmeistä, että keksinnön suoritusmuodoista riippuen kokoonpano voi olla erilainenkin. Laitteisto voi sijaita PC:llä (personal computer), verkon palvelimella tai laitteiston eri osat voivat sijaita fyysisesti eri paikoissa, kunhan yhteydet niiden välillä ovat riittävän nopeat.

Kuvio 2 esittää erään edullisen suoritusmuodon mukaista käsiteltävää syötetietovirran osaa 200, joka siis tyypillisesti tallennetaan keskusmuistiin käsittelyn ajaksi. Syötetietovirta on tässä suoritusmuodossa luonnollista kieltä ja syötetietovirran kerralla käsiteltävä osa 200 on tyypillisesti lause tai virke. Tämä käsiteltävä osa 200 on jaettu elementteihin 211, 212, 213, 221, 222, 223, jotka ovat luonnollisten kielten tapauksessa yleensä sanoja mahdollisine etu- ja/tai jälkiliitteineen. Sanaa edeltävä määräinen tai epämääräinen artikkeli kuuluu tyypillisesti samaan elementtiin itse sanan kanssa.

Käsiteltävän syötetietovirran osan 200 elementit 211, 212, 213, 221, 222, 223 on kuviossa 2 jaoteltu kahteen segmenttiin 210, 220. Tässä tapauksessa segmentointi

on tehty tunnistamalla "vaikka"-elementti, joka nyt kuuluu sellaisten sanojen listaan, jotka aloittavat uuden segmentin. Vastaavia listoja esiintyy yleisesti luonnollista kieltä käsittelevässä kirjallisuudessa. Segmentit voivat koostua yhdestä tai, kuten kuvassa on esitetty, useammasta elementistä. Segmentointi tehdään tiettyjen edullisesti massamuistissa olevien sääntöjen perusteella, jotka voivat perustua esimerkiksi tiettyihin helposti tunnistettaviin sanoihin tai käsiteltävän syötetietovirran osan ja tietämyskannan sisällön vastaavuuteen. Eräitä käyttökelpoisia segmentointisääntöjä on esitelty tarkemmin esimerkiksi patenttijulkaisussa FI 103156. Esimerkiksi suomen kielelle voidaan käyttää erinäisiä segmentointisääntöjä. Eräs tyypillinen ratkaisu on, että segmentiksi valitaan pisin vastaava segmentti tietämyskannasta tai fraasisanakirjasta. Kun mahdollisimman paljon elementtejä käsitellään yhdellä kertaa, luokittelu tehostuu ja kääntämiseen liittyvät segmenttien yhdistelyn ja käännösten muokkaamisen ongelmat voidaan välttää paremmin. Usein segmentti katkaistaan välimerkkiin tai sanaan, joka aloittaa sivulauseen tai fraasin. Segmentointi voidaan tehdä myös käyttäjän ohjeiden ja valintojen mukaisesti. Lisäksi segmentti voidaan rajata tekstityypin tai ominaisuuksien perusteella, esimerkiksi siten, että lihavoidut peräkkäiset sanat käsitellään yhtenä segmenttinä. Myös usean tunnistamattoman elementin jono voidaan valita yhdeksi segmentiksi.

On selvää, että segmentointisäännöt ovat kielikohtaisia ja vaihtelevat jonkin verran kielittäin. Yleisenä, lähes kaikkiin luonnollisiin kieliin soveltuvana sääntönä voidaan pitää sitä, että valitaan segmentiksi jokin jo tietämyskannassa oleva segmentti. Lisäksi jos käsiteltävän syötetietovirran keskellä tai lopussa oleva segmentti tunnistetaan jonkun säännön perusteella, sitä edeltävää elementtijonoa ja sitä seuraavaa elementtijonoa voidaan käsitellä erillisinä segmentteinä. Formaalien kielien tapauksessa elementit ovat tyypillisesti merkkijonoja tai yksittäisiä käskyjä. Segmentit voidaan erotella esimerkiksi koostuviksi käskyistä ja niiden parametreista tai segmentti voi päättyä rivinvaihtoon tai muuhun käytettyyn merkkiin, merkkijonoon tai erikoismerkkiin.

Kuviossa 3 on esitetty osa keksinnön erään edullisen suoritusmuodon mukaisesta tietämyskannasta. Tietämyskannassa on kaksi tallennettua segmenttiä: segmentti 31, joka sisältää elementit 311, 312, 313, ja segmentti 32, joka sisältää elementit 321, 322, 323. Segmentin 32 elementit 321, 322, 323 on analysoitu ja niiden analyysin tulokset on merkitty elementin alle. Tässä luonnollisen kielen esimerkkitapauksessa elementti 321 "kissa" on analyysin perusteella substantiivi (noun), yksikkö (sg, singular), nominatiivi (nom). Elementti 322 "kävelee" on analysoitu verbiksi (verb) yksikön kolmannessa persoonassa (sg 3). Elementti 323 "katolla" on substantiivin

(noun) yksikön (sg) adessiivi (ades). Luonnolliselle kielelle on tässä tehty leksikaalinen (sanastollinen) tai morfologinen (muoto-opillinen) analysointi jollain tunnetulla tehokkaalla menetelmällä. Tämän menetelmän etuna on se, että käännösvastineen tuottaminen sanoille, joita ei ennestään löydy tietämyskannasta, onnistuu hyvin näiden elementeille annettavien morfologisten leimojen perusteella. Vaihtoehtoisesti voidaan käyttää esimerkiksi syntaktisia (lauseopillisia, syntaksiin perustuvia) tai semanttisia (merkitysopillisia) sääntöjä. Formaalien kielten tapauksessa säännöt voivat perustua esimerkiksi kielen formaaliin esitystapaan ja matriisielementtejä käsiteltäessä analyysi voi perustua matriisin normiin, matriisin esittämän kuvan valoisuuteen tai matriisia esittävän kosinimuunnoksen kolmeen ensimmäiseen kertoimeen. Vaikka keksinnön mukaisesti elementeille tuotetaan tietyt analyysitulokset, mitään jäsennyspuita ei muodosteta.

Kuvion 3 segmentti 33 on tietämyskannan yksi vastinsegmentti. Tässä on kuvattu vastinsegmentti tietämyskannan segmentille 32. Näiden segmenttien 32 ja 33 vastaavuustiedon perusteella elementtiä 321 vastaa elementti 331, elementtiä 322 vastaa elementti 332 ja elementtiä 323 vastaa elementti 333. Vastinelementtien analyysitulokset eivät välttämättä ole samat eri kielissä eikä myöskään niiden järjestys tai lukumäärä. Tyypillisesti vastinsegmentti tai segmenttien välinen assosiaatiotieto sisältää järjestystiedon, joka kertoo, missä sanajärjestyksessä, tai yleisemmin elementtijärjestyksessä, vastaavan segmentin elementit voivat olla. Tätä järjestystietoa ei ole esitetty kuviossa 3. Vastinsegmenttejä voi olla useampiakin, myös yhdellä kieliparilla. Tällöin vastinsegmenteistä yleensä yksi on optimaalisin vastinsegmentti, mikä tarkoittaa voi esimerkiksi yleisintä, käytetyintä tai asiayhteydessä suositeltavinta vastinsegmenttiä. Muitakin vaihtoehtoisia vastinsegmenttejä voidaan käännöstä muodostettaessa käyttää. Kun vastinsegmenttejä on useampia, assosiaatiotiedon on lisäksi sisällettävä tieto siitä, mihin vastinsegmenttiin mikäkin järjestystieto kohdistuu. Esimerkiksi suomenkielisessä segmentissä englanninkieliseen vastinsegmenttiin viittaava assosiaatiotieto voi sisältää järjestystiedon, jonka mukaan suomenkielisen segmentin ensimmäistä elementtiä vastaa englanninkielisessä ensimmäinen elementti, toista kolmas ja kolmatta toinen elementti. Vastaavan suomenkielisen segmentin saksankieliseen vastinsegmenttiin viittaava järjestystieto voi olla sellainen, että ensimmäiselle suomenkielen elementille ei ole lainkaan vastinetta, toista vastaa neljäs saksankielinen elementti, kolmatta kolmas ja näiden lisäksi vastinsegmentissä on kaksi muuta elementtiä sen alussa. Formaaleja kieliä käsiteltäessä järjestystieto on oleellinen ja on tärkeää assosioida kielten toiminnallisesti toisiaan vastaavat osiot toisiinsa.

Tarkastellaan kuviossa 2 esitetyn syötetietovirran 200 ensimmäisen käsiteltävän osan eli segmentin 210 "koira kävelee kadulla" kääntämistä englanninkieliseksi kuviossa 3 esitetyn tietämyskannan avulla keksinnön edullisen suoritusmuodon mukaisesti. Aluksi syötetietovirran 200 segmenttejä verrataan tietämyskannan segmentteihin. Esimerkkinä olevassa tapauksessa elementit ovat luonnollisen kielen sanoja, joita käsitellään tässä vertailussa segmentin kokoisina yhtenäisinä elementtijonoina. Tällainen jono voidaan muodostaa eri tavoin, kuten esimerkiksi vain yhdistämällä segmentin elementit toisiinsa tai laittamalla elementtien väliin jokin ennalta sovittu merkki. Keksinnön kannalta on oleellista, että syötetietovirran segmentti on verrattavissa tehokkaasti tietämyskannan segmenttiin, eli segmentit ovat saman muotoisia. Tehokkaaseen vertailuun voidaan käyttää esimerkiksi tunnettuja indeksointitekniikoita tai tiedonhallintajärjestelmien tarjoamia indeksointi- ja levynkäsittelyn optimointimekanismeja.

Tietämyskannan ensimmäinen segmentti 31 ei vastaa syötetietovirran 200 segmenttiä 210. Näillä segmenteillä on sama ensimmäinen elementti 211, 311, mutta tässä vertailu tehdään segmentille kokonaisuutena. Tietämyskannan toinenkaan segmentti 32 ei vastaa syötetietovirran 200 segmenttiä 210, vaikka näidenkin segmenttien toiset elementit, 212 ja 322, ovat samat. Syötetietovirran segmentin vertailua tietämyskannan segmentteihin voidaan tehostaa käyttämällä tunnettuja indeksointi- ja hakumenetelmiä. Mikäli elementeiltään täysin vastaavaa segmenttiä ei tietämyskannasta löydy, syötetietovirran 200 segmentin 210 elementit 211, 212, 213 analysoidaan ja jokaiselle elementille saadaan jokin analyysitulos. Tämän jälkeen tarkastellaan edelleen segmenttiä luokiteltuna kokonaisuutena. Nyt tutkitaan yhtenäistä segmentin pituista, sovitulla tavalla muodostettua jonoa analyysituloksia eli segmentin luokitusta ja verrataan sitä tietämyskannan vastaaviin analyysitulosjonoihin eli luokitteluihin. Tämän vertailun tuloksena syötetietovirran 200 segmenttiä 210 vastaa tietämyskannassa segmentti 32. Tietämyskannan segmentille 32 haetaan vastinsegmentti 33 tietämyskannasta ja analyysitulosten perusteella löydetyn tietämyskannan segmentin 32 elementtejä 321, 322, 323 verrataan syötetietovirran 200 vastaaviin elementteihin 211, 212, 213. Näistä elementeistä toisiaan täysin vastaavat keskimmäiset, eli tulostietovirta koostuu elementeistä, joista keskimmäiselle löytyy vastinelementti. Syötetietovirran ensimmäiselle ja viimeiselle elementille muodostetaan tulostietovirtaan vastinelementit esimerkiksi hakemalla syötetietovirran elementille vastinelementti elementtien ja vastinelementtien tietokannasta ja generoimalla tarkka vastinelementin analyysituloksen mukainen elementtimuoto erillisellä generointiohjelmalla. Suoritusmuodosta riippuen edellä esitetyt käännösvaiheet voidaan suorittaa kullekin käsiteltävän syötetietovirran osan segmentille alusta loppuun tai koko

käsiteltävälle syötetietovirran osalle kukin vaihe segmentti kerrallaan. Edellä esitetyssä suoritusmuodossa edellä esitetyt käännösvaiheet suoritetaan seuraavaksi kuvion 2 toiselle segmentille 220.

Edullisen suoritusmuodon mukainen tulostietovirran osa on esitetty kuviossa 4. Kuviossa 4 on löydetty syötetietovirtaa vastaava segmentti luokittelun perusteella ja syötetietovirran elementille on löydetty tietämyskannasta vastinelementti 402. Elementeille 401 ja 403 löydettiin tietämyskannasta vastaava analyysitulos, jonka perusteella kyseisistä runkosanoista, substantiiveista ei ole tietoa, mutta muoto on sama kuin vastinelementtien analyysituloksissa määritetty. Tämä tarkoittaa sitä, että sanan liitteet eli pre- ja postpositiot ovat samat kuin analyysitulosta vastaavalla muodolla. Tyypillisesti tämä puuttuva osa kysytään käyttäjältä, mutta se voidaan myös esimerkiksi hakea jostain sähköisestä sanakirjasta. Kuviossa 3 esitetty segmenttien tietämyskanta ja vastinsegmenttien tietämyskanta ovat keskenään symmetriset, joten niitä voidaan käyttää kaksisuuntaisesti, eli syötetietovirta voikin olla vastinsegmenttien muotoista ja tulostietovirta tietämyskannan segmenttien muotoista. Vastaava kaksisuuntaisuus voidaan toteuttaa myös useamman kielen kesken sekä rinnakkaisesti että sarjamuotoisesti. Rinnakkaiset kielet ovat tasa-arvoisia ja käännöksen syöte- ja kohdekielet voidaan valita näistä. Sarjamuotoisessa järjestelyssä esimerkiksi kolmas kieli voi toimia niin sanottuna välikielenä, jonka kautta käännös kahden muun kielen välillä aina tehdään.

Kuviossa 5 on esitetty erään edullisen suoritusmuodon mukainen menetelmä datan luokittelemiseksi. Lohkossa 501 luetaan syötetietovirrasta kerralla käsiteltävä osa, joka esimerkiksi luonnollista kieltä luokiteltaessa voi olla esimerkiksi tiedonhakupyyntö, lause, virke tai käsky parametreineen. Käsiteltävästä syötetietovirran osasta erotellaan elementit, jotka tässä käsiteltävän esimerkin mukaisesti ovat siis sanoja liitteineen tai merkkijonoja. Lohkossa 502 käsiteltävä syötetietovirran osa ryhmitellään segmentteihin tiettyjen muistiyksikköön tallennettujen sääntöjen tai käyttäjän määritysten mukaisesti. Segmentti voi sisältää yhden tai useamman elementin. Vaiheessa 503 verrataan yhden tai useamman elementin sisältäviä syötetietovirran segmenttejä kokonaisuutena tietämyskannassa jo oleviin segmentteihin. Mikäli sisällöltään täysin vastaavaa segmenttiä ei löydy, siirrytään lohkoon 504, jossa elementit analysoidaan joko jollain järjestelmän sisäisellä mekanismilla tai jollain erillisellä analysaattorilla. Jokaisesta elementistä tuotetaan analyysitulos, joka tyypillisesti luonnollisen kielen tapauksessa perustuu leksikaaliseen tai morfologiseen analyysiin, formaalin kielen tapauksessa syntaktiseen analyysiin.

Vaiheessa 505 verrataan segmenteittäin syötetietovirran elementtien analyysituloksia, eli segmenttien luokitusta, tietämyskantaan tallennettujen segmenttien luokituksiin. Jollei vastaavaa segmenttiä luokittelun perusteellakaan löydy, suoritetaan poikkeuskäsittely lohkossa 506. Poikkeuskäsittely on jokin ennalta määrätty toiminto tai menettely, jossa voidaan esimerkiksi luoda syötetietovirran segmentistä uusi tietämyskantasegmentti, käsitellä jokaista elementtiä yhtenä segmenttinä tai suorittaa uusi segmentointi. Tämän jälkeen suoritus siirtyy vaiheeseen 508. Jos vaiheessa 505 verratut analyysitulokset vastaavat toisiaan, siirtyy suoritus lohkoon 507, jonne siirrytään myös vaiheesta 503, jos syötetietovirran ja tietämyskannan segmentit vastaavat toisiaan. Lohkossa 507 assosioidaan syötetietovirran segmenttiin sitä vastaava tietämyskannassa jo oleva segmentti.

Vaiheessa 508 tarkastetaan, onko käsiteltävässä syötetietovirran osassa vielä käsittelemättömiä segmenttejä. Jos segmenttejä on vielä käsittelemättä, siirtyy suoritus alkuun lohkoon 503, jotta kaikki käsiteltävän syötetietovirran osan sisältämät segmentit käydään läpi. Muuten siirrytään lohkoon 509 tarkastelemaan, sisältyvätkö nyt luokitellut segmentit johonkin ylemmän tason segmenttiin. Tällainen tilanne voi esiintyä esimerkiksi, kun keksinnön edullisen suoritusmuodon mukaista luokittelijaa käytetään luonnollisia tai formaalia kieliä käännettäessä tai valuuttoja konvertoitaessa. Ylemmän tason segmentit selkeyttävät ja yksinkertaistavat toimintaa esimerkiksi silloin, kun valuuttojen tunnukset siirtyvät useampia numeerisia elementtejä sisältävien rakenteiden yli eri kielien välillä, formaalissa kielessä on sisäkkäisiä silmukkarakenteita, tai kun luonnollinen kieli on saksa ja segmentti sisältää saksankielisen lauseen, jonka rakenne ei vastaa vastinkielen rakennetta. Saksankielen esimerkkitapauksessa ylemmäksi tasoksi voi muodostua segmentti, jonka ensimmäinen alisegmentti sisältää tietyn konjunktion, toinen tietyn luokituksen mukaisia segmenttejä, jotka sisältävät useita tuntemattomia elementtejä ja viimeinen alisegmentti verbiksi luokitellun elementin. Näin voidaan yleistää useita samankaltaisia tilanteita ja muodostaa niitä kuvaava geneerinen segmentti tietämyskannan ylemmälle tasolle välittämättä siitä, mitä tarkalleen ottaen lauseen elementit ovat. Tämä pienentää edelleen tietämyskannan kokoa ja nopeuttaa vertailuja.

Lohkossa 510 tarkastellaan useamman segmentin muodostamaa jonoa ja tutkitaan, kuuluvatko tai täsmäävätkö edellä käsitellyt segmentit tai segmenttien jono johonkin hierarkkisesti ylemmän tason segmenttiin. Ylemmän tason segmentti voi koostua yhdestä tai useammasta alemman tason segmentistä. Jos ylempiä segmenttejä löytyy, myös niille haetaan luokitustulos 511 vastaavasti kuin alemman tason segmenteillekin. Jos vastaavaa ylemmän tason segmenttiä ei tietämyskannasta löydy,

jää luokitteluksi alisegmenttien jono. Jos ylemmän tason segmenttejä ei oltu muodostettu tai kun luokittelu lohkossa 511 on tehty, tarkastellaan lohkossa 512, onko käsiteltävässä syötetietovirran osassa vielä segmenttejä, jotka voidaan assosioida joksikin toiseksi ylemmän tason segmentiksi. Mikäli tällaisia löytyy, suoritusta jatketaan lohkosta 510. Kun segmenteistä muodostuvia ylemmän tason segmenttejä ei enää löydetä, tutkitaan vielä vaiheessa 513 muodostavatko löydetyt ylemmän tason segmentit edelleen kolmannen tason segmenttejä. Jos vielä ylemmän tason segmenttejä löytyy, jatketaan suoritusta lohkosta 509. Tyypillisesti alimman tason segmentit sisältävät elementtejä, seuraavan ylemmän tason segmentit sisältävät segmenttejä ja mahdollisesti myös elementtejä. Mitä ylemmälle segmenttitasolle mennään, sitä enemmän luonnollisten kielten segmentit sisältävät tiettyjä sopimuksellisia vakioehtoja, kuten esimerkiksi tekstikappaleen kontekstin. Formaalien kielten tapauksessa segmentit voivat olla esimerkiksi käskyjä parametreineen tai kielen lauseita, jotka siis erotellaan toisistaan tyypillisesti jonkin merkin avulla. Tällöin ylemmän tason segmentti voi sisältää rakenteellista tietoa, esimerkiksi tiedon silmukasta, sisäkkäisistä silmukoista tai aliohjelmista. Mitä ylemmälle segmenttitasolle mennään, sitä enemmän formaalien kielten segmenttien sisältö lähestyy algoritmikuvausta.

Kun hierarkkiset segmentit on läpikäyty ja luokiteltu, lohkossa 514 raportoidaan käsitellyn syötetietovirran osan luokitus yhden tai useamman ylemmän tason hierarkkisten segmenttien jonona. Kuviossa 5 esitetyn menetelmän mukainen datan luokittelija siis assosioi käsiteltävään syötetietovirran osaan jonon mahdollisesti hierarkkisia tietämyskannassa olevia segmenttejä. Kun käsitellään hierarkkisia rakenteita, hierarkkisten alisegmenttien järjestystieto on tyypillisesti ylemmän tason segmentissä. Tämä järjestystieto määrittää alemman tason segmenttien järjestyksen eli esimerkiksi luonnollisen kielen tapauksessa sanajärjestyksen, formaalin kielen tapauksessa käskyn tai aliohjelmakutsun parametrit, niiden tyypin, lukumäärän ja järjestyksen.

Kuvion 6 suoritusmuodossa on esitetty uusien segmenttien ja vastinsegmenttien tuottamista tietämyskantoihin oppimisen avulla eli tietämyskannan kasvattamista ilman käyttäjän vuorovaikutusta. Vaiheessa 601 luetaan kaksi toisiaan vastaavaa syötetietovirran osaa. Kuvion 6 mukaisen menetelmän suorittaminen edellyttää, että käytettävissä on kaksiosainen syötetietovirta, jonka tiedetään sisältävän sama data kahdessa eri esitysmuodossa, jotka ovat toistensa täydellisiä vastineita. Lohkossa 602 luokitellaan luetut toisiaan vastaavat syötetietovirran osat esimerkiksi sillä luokittelumenetelmällä, joka on esitetty kuvion 5 suoritusmuodossa. Lohkossa 603 tallennetaan kumpikin syötetietovirran osa tietämyskantaan ja tallennetuille syötetie-

tovirran osille luodaan vastaavuustieto tietämyskannan avulla siten, että etsitään tietämyskannassa jo olevia segmenttejä vastaavia osia sekä luokittelutulosten vastaavuuksia. Tässä esitettyjä tyypillisiä uutta syötetietovirtaa segmentoitaessa käytettäviä vertailukriteerejä voidaan käyttää useissa muissakin keksinnön edullisissa suoritusmuodoissa. Ensisijainen valinta on sellainen segmentti, joka löytyy tietämyskannasta ja jonka jokaista elementtiä vastaa juuri sama syötetietovirran elementti. Tällöin valitaan pisin mahdollinen tietämyskannan vastaava segmentti ja assosioidaan se tarkasteltavaan syötetietovirran osaan. Seuraavaksi tarkastellaan analyysituloksia. Jos useammalla tietämyskannan segmentillä on syötetietovirran tarkasteltavaa osaa vastaava analyysitulos, valitaan se, jonka mahdollisimman usea elementti on vastaava kuin tarkasteltavan syötetietovirran osan. Jos vastaavia elementtejäkin on useammalla tietämyskannan segmentillä saman verran, valitaan kulloinkin tilanteeseen ja sovellukseen sopivin toiminto, joka voi olla esimerkiksi se, että segmentti valitaan käyttötiheyden mukaan siten, että valitaan se, jota on käytetty useimmin. Segmentillä voi myös olla jokin semantiikkaluokitus, eli esimerkiksi toimialamääritys, joka määrittää segmentin kuuluvan tiettyyn alaan, kuten paperiteknologiaan tai biotekniikkaan. Lisäksi kullakin elementillä voi olla vastaava semanttinen luokitus. Segmentit voivat lisäksi sisältää niin sanotun leiman, eli prioriteetin, joka kertoo vaikkapa, että tietty segmentti on virallinen käännös tai tiettyä segmenttiä ei pidä käyttää käännöksen tulostietovirran segmenttinä, vaan ainoastaan syötetietovirran luokitusta tehtäessä.

Lohkossa 604 testataan, oliko jompikumpi käsiteltävistä syötetietovirran osista kokonaisuutena jo tietämyskannassa. Jos syötetietovirran osaa vastaava lohko löytyy tietämyskannasta, on tietämyskannassa myös tieto tällaisen syötetietovirran osan sisältämistä segmenteistä. Löydetyn segmenttijaon mukaisesti lohkossa 605 myös syötetietovirran osa jaetaan segmentteihin. Lisäksi lohkossa 605 haetaan käännökset eli vastinsegmentit ja niiden vastaavuustieto etsimällä tietämyskannasta tunnettujen segmenttien ja luokitusten vastaavuuksia, minkä jälkeen suoritus loppuu lohkossa 610. Jos lohkossa 604 ei löydy koko syötetietovirran osaa vastaavaa lohkoa tietämyskannasta, käsittely siirtyy lohkoon 606.

Lohkossa 606 vielä käsittelemättömiä syötetietovirran osia verrataan tietämyskannan segmentteihin millä hyvänsä sopivalla segmenttikoolla ja tietämyskannasta etsitään parhaiten käsittelemätöntä syötetietovirran osaa vastaavaa segmenttiä. Jos tietämyskannasta löydetään segmentti, joka vastaa jotain osaa käsiteltävästä syötetietovirran osasta, haetaan lohkossa 608 tälle syötetietovirran osalle eli segmentille tietämyskannasta vastaava segmentti ja vastaavuustieto. Näiden perusteella varsinai-

nen käännös eli vastinsegmentti löytyy tietämyskannasta. Lohkossa 609 tarkastetaan, onko käsiteltävästä syötetietovirran osasta vielä osioita käsittelemättä. Tästä siirrytään lohkoon 606 käsittelemään loppua syötetietovirran osaa, kunnes kaikille syötetietovirran segmenteille on luotu tai löydetty vastaavat segmentit. Jos lohkossa 606 ei löydetä tarpeeksi hyvää segmenttiä tietämyskannan kummastakaan osasta, siirrytään lohkoon 607. Vaiheessa 607 jäljelle jääneitä syötetietovirran osia sovitetaan toisiinsa, ja niistä tuotetaan segmentit ja luodaan vastinsegmenttitieto. Tämän jälkeen lopetetaan suoritus lohkossa 610.

Varsinainen datan kääntäminen automaattisesti tapahtuu keksinnön erään edullisen suoritusmuodon mukaan kuviossa 7 esitetyllä tavalla. Aluksi luetaan syötetietovirran osa lohkossa 701. Käsiteltävä syötetietovirran osa myös luokitellaan lohkossa 701, mahdollisesti hierarkkisten segmenttien jonoksi, esimerkiksi kuvion 5 yhteydessä esitetyn luokittelumenetelmän mukaisesti. Lohkossa 702 jokaiselle käsiteltävän syötetietovirran osan segmentille haetaan vastinsegmentti vastinsegmenttien tietämyskannasta. Jotkut segmenteistä voivat muodostaa myös ylemmän tason segmentin. Seuraavaksi haetaan vastinsegmenttejä löydetyille ylemmän tason segmenteille tietämyskannasta lohkossa 703. Jos ylemmän tason segmenteille ei löydetä vastinsegmenttejä, jää tuloksèksi jono alemman tason segmenttejä. Vastinsegmentit ja edelleen vastinsegmenttien elementit järjestetään järjestystiedon mukaiseen järjestykseen. Järjestystietohan voi sijaita segmenteissä tai assosiaatiotiedossa eli tietämyskannan segmentit vastinsegmentteihinsä yhdistävässä vastaavuustiedossa. Tämä vastaavuustieto puolestaan voi sijaita joko segmenteissä tai niistä erillään. Sellaisille elementeille, joille ei ole vielä löydetty vastinelementtejä, tuotetaan vastinelementit lohkossa 704. Näitä vastinelementtejä voidaan hakea vastinelementtien tietokannasta tai tuottaa analyysitulosten perusteella jollain sopivalla generaattorilla. Generaattori voi käyttää hyväkseen esimerkiksi sanakirjatyyppistä vastinelementtien tietokantaa vastinelementin rungon hakemiseksi ja muokata sen analyysitulosten mukaisesti haluttuun muotoon. Lopuksi lohkossa 705 tuotetaan käsiteltävän syötetietovirran osaa vastaava tulosvirran osa vastinsegmenttien sisältämien elementtien sekä generoitujen vastinelementtien jonona, jotka on järjestetty järjestystiedon mukaisesti segmenttien sisällä. Kun käännös on valmis, se voidaan vielä lisätä tietämyskantaan.

Usein kuitenkin tietämyskannan koko halutaan pitää suhteellisen pienenä, koska haku on tällöin nopeampaa, eikä tietorakenne vie paljoa tilaa, vaan mahtuu keskusmuistiin. Varsinkin hierarkkisia segmenttejä sisältäviin tietämyskantoihin on turha

tallentaa kaikkia sisältövaihtoehtoja, koska ne löytyvät olemassa olevien tietojen perusteella tehokkaammin kuin isosta tietämyskannasta hakemalla.

Tässä hakemuksessa käsitellään esimerkkitapauksena luonnollisen kielen kääntämistä, mutta on ilmeistä, että keksinnön mukaista menetelmää voidaan yhtä hyvin soveltaa esimerkiksi puheen, kuvien ja formaalien kielien luokitteluun ja tunnistamiseen. Lisäksi käsiteltävät elementit voivat olla esimerkiksi lukuja, matriiseja, merkkijonoja, konekielisiä käskyjä tai parametreja. Formaalien kielten kääntäminen ja luokittelu on erittäin tärkeää, kun halutaan käyttää ja yhtenäistää erimuotoista tietoa ja dataa eri lähteistä.

Yleensäkin haettaessa tietoja ja tehtäessä kyselyjä on tärkeää, että tunnistetaan ja otetaan osaksi tulostietovirtaa myös läheisiksi tulkittavat, löydetyt segmentit. Tällöin kriteereinä voidaan käyttää esimerkiksi jo tässä hakemuksessa mainittua semanttista läheisyyttä, jossa tutkitaan merkityksiä. Sovellusmuodosta riippuen voi olla edullista tarkastella vaihtoehtoisesti tai lisäksi vaikkapa leksikaalista eli sanastollista tulkintaa, morfologista eli muoto-opillista tulkintaa tai syntaktista eli lauseopillista tai syntaksiin liittyvää tulkintaa. Mikäli toivottua luokittelua tai käännöstä ei saada tuotettua, voidaan keksinnön erään edullisen suoritusmuodon mukaan suorittaa esimerkiksi luokittelu tai jokin muu osatoiminto tai koko käännös käyttäen vastaavaa keksinnön edullisen suoritusmuodon mukaista laitteistoa ja menetelmää, johon on olemassa tai voidaan muodostaa tietoliikenneyhteys. Toinen vastaava järjestelmä voi esimerkiksi käsitellä ensisijaisesti tietyn erityisalan segmenttejä tai elementtejä. Lisäksi useamman laitteiston käytössä voi olla yhteen muistiyksikköön tallennettuna esimerkiksi segmentointisääntöjä, poikkeussääntöjä ja muunnossääntöjä sekä listauksia semanttisesti, leksikaalisesti, morfologisesti ja syntaktisesti toisiaan vastaavista elementeistä ja segmenteistä.

L 2

**Patenttivaatimukset**

1. Menetelmä elementtejä (211, 212, 213, 221, 222, 223) sisältävän syötetietovirran (200) datan käsittelemiseksi segmenttejä sisältävän tietämyskannan avulla, **tunnettu** siitä, että menetelmä sisältää vaiheet, joissa

- luetaan (501) käsiteltävä osa syötetietovirrasta (200) ja jaetaan käsiteltävä syötetietovirran osa elementteihin (211, 212, 213, 221, 222, 223),

- ryhmitellään käsiteltävä osa syötetietovirtaa (200) segmenteiksi (502), joista jokainen segmentti (210, 220) sisältää yhden tai useampia elementtejä (211, 212, 213, 221, 222, 223),

- analysoidaan käsiteltävän syötetietovirran osan elementit ja tuotetaan analyysitulosten perusteella segmenttikohtainen luokitus,

- verrataan syötetietovirran segmenttien (210, 220) luokitusta tietämyskannan segmenttien (31, 32) luokituksiin ja assosioidaan tietämyskannan segmentti sen luokitusta vastaavaan syötetietovirran segmenttiin, ja

- raportoidaan tulos, joka on käsiteltävään syötetietovirran osaan assosioitu joukko tietämyskannassa olevia segmenttejä.

2. Patenttivaatimuksen 1 mukainen menetelmä, **tunnettu** siitä, että ainakin yksi segmentti (210, 220) sisältää ainakin kaksi elementtiä (211, 212, 213, 221, 222, 223), ja segmenttikohtainen luokitus määritetään ainakin kahden mainitun elementin (211, 212, 213, 221, 222, 223) analyysituloksen perusteella.

3. Patenttivaatimuksen 1 mukainen menetelmä, **tunnettu** siitä, että elementtien analyysitulokset katenoidaan segmenttikohtaisen luokituksen muodostamiseksi.

4. Patenttivaatimuksen 1 mukainen menetelmä, **tunnettu** siitä, että syötetietovirran segmentin luokitus toimii hakuavaimena etsittäessä samoin luokiteltua tietämyskannan segmenttiä.

5. Patenttivaatimuksen 1 mukainen menetelmä, **tunnettu** siitä, että segmenteiksi ryhmittelyn jälkeen tehdään vaihe, jossa käsiteltävää syötetietovirran osaa verrataan segmenteittäin (210, 220) tietämyskannan segmentteihin (31, 32) ja toisiaan vastaa-

vat segmentit assosioidaan keskenään, minkä jälkeen analysointivaihe tehdään ainoastaan niille segmenteille, joille ei löydy vastaavaa tietämyskannan segmenttiä.

6. Patenttivaatimuksen 5 mukainen menetelmä, **tunnettu** siitä, että jos yhtä syötetietovirran segmenttiä vastaa tietämyskannan segmentteihin verrattaessa useampi segmentti, valitaan niistä yksi segmentti soveltaen ainakin yhtä seuraavista kriteereistä:

- valitaan segmentti, jossa on eniten syötetietovirran elementtejä,
- valitaan segmentti, jonka käyttäjä ilmaisee,
- valitaan segmentti, jota on käytetty useimmin,
- valitaan segmentti, jonka semanttinen luokitus vastaa syötetietovirran vastaavan osan luokitusta,
- valitaan segmentti, jonka elementtien semanttinen luokitus vastaa syötetietovirran vastaavan osan luokitusta.

7. Patenttivaatimuksen 1 mukainen menetelmä, **tunnettu** siitä, että tietämyskantaan sisällytetään eri pituisia, osittain samansisältöisiä segmenttejä, joiden avulla käsiteltävä osa syötetietovirtaa ryhmitellään segmenteiksi optimaalisesti tapauskohtaisesti.

8. Patenttivaatimuksen 1 mukainen menetelmä, **tunnettu** siitä, että syötetietovirran ryhmittely segmenteiksi tehdään ainakin jollain seuraavista menetelmistä:

- segmentiksi valitaan jo tietämyskannassa oleva, syötetietovirran osaa elementeiltään tai luokitukseltaan vastaava segmentti,
- segmentti määritetään käyttäjän ohjeiden mukaisesti,
- kielellisestä kokonaisuudesta muodostetaan segmentti,
- fraasista muodostetaan segmentti,
- segmentti katkaistaan välimerkkiin,
- segmentti katkaistaan tiettyihin listattuihin välikesanoihin,
- segmentti muodostetaan jäljelle jääneestä syötetietovirran osasta, kun syötetietovirran osasta muilla keinoilla löydetyt segmentit on poistettu.

9. Patenttivaatimuksen 1 mukainen menetelmä, **tunnettu** siitä, että segmentit muodostavat hierarkkisia rakenteita, joissa tietty ylemmän tason segmentti sisältää tietoa tietyistä alemman tason segmenteistä, ja menetelmä sisältää vaiheen, jossa käsiteltävään syötetietovirran osaan (200) assosioidaan tietämyskannan ylemmän tason segmenttejä (509), jotka sisältävät syötetietovirran segmentteihin assosioituja tietämyskannan alemman tason segmenttejä.

10. Patenttivaatimuksen 1 mukainen menetelmä, **tunnettu** siitä, että syötetietovirran segmentille suoritetaan poikkeuskäsittely (506) tiettyjen ohjeiden mukaisesti tilanteessa, jossa vastaavaa segmentin luokitusta ei löydy tietämyskannasta.

11. Patenttivaatimuksen 1 mukainen menetelmä, **tunnettu** siitä, että elementeille tehtävä analyysi on morfologisen analyysi, jonka tuloksena tuotetaan tiettyjä, mainittuja elementtejä kuvaavia piirteitä.

12. Patenttivaatimuksen 1 mukainen menetelmä, **tunnettu** siitä, että datan kääntämiseksi kohdekielelle haetaan tuloksen segmenteille (210, 220) vastinsegmentit (33) kahden tai useamman kielen tietämyskannasta, ja tuotetaan tulosvirtana vastinelementtejä (401, 402, 403) sisältävä vastinsegmenttien (400) joukko.

13. Patenttivaatimuksen 12 mukainen menetelmä, **tunnettu** siitä, että syötetietovirran elementeille (211, 212, 213, 221, 222, 223), joille ei löytynyt vastaavuuksia tietämyskannasta, tuotetaan vastinelementit tiettyjen, tietämyskannan elementteihin (331, 332, 333) liittyvien analyysitulosten perusteella ja/tai erillisen, elementtejä tuottavan generaattorin avulla.

14. Patenttivaatimuksen 12 mukainen menetelmä, **tunnettu** siitä, että datan kääntämisessä tuotettava tulosvirta sisältää vastinsegmenttien (400) elementtejä (401, 402, 403) ja erikseen tuotettuja elementtejä segmenttijonona siten, että kunkin segmentin sisäinen vastinelementtien järjestys määritetään vastinsegmenttien sisältämän järjestystiedon perusteella.

15. Patenttivaatimuksen 12 mukainen menetelmä, **tunnettu** siitä, että datan kääntämisessä tuotettava tulosvirta sisältää vastinsegmenttien (400) elementtejä (401, 402, 403) ja erikseen tuotettuja elementtejä segmenttijonona siten, että kunkin segmentin sisäinen vastinelementtien järjestystieto määritetään segmenttien ja niiden vastinsegmenttien välisessä vastaavuustiedossa.

16. Patenttivaatimuksen 1 mukainen menetelmä, **tunnettu** siitä, että tietämyskannan muodostamiseksi

- luetaan kaksi toisiaan vastaavaa syötetietovirran osaa (601) ja jaetaan ne elementteihin,

- luokitellaan kerralla käsiteltävät syötetietovirtojen osat,

- haetaan käsiteltävälle syötetietovirran osalle segmenttijako, vastinsegmentit ja edellisten väliset vastaavuustiedot (603, 605, 608) tietämyskannassa olevien segmenttien ja niiden luokituksen perusteella, ja

- sovitetaan segmentoimattomat, vastinsegmentittömät osat käsiteltävistä syötetietovirroista toisiinsa (607), muodostetaan niistä segmentit, luodaan segmenteille vastinsegmentit ja niiden välinen vastaavuustieto.

17. Patenttivaatimuksen 16 mukainen menetelmä, **tunnettu** siitä, että segmenttien vastaavuustieto, vastinsegmentit ja segmenttijako luodaan tietämyskantaan (33) jo tallennettujen segmenttien ja/tai niiden luokittelun perusteella.

18. Laitteisto elementtejä (211, 212, 213, 221, 222, 223) sisältävän syötetietovirran (200) datan käsittelemiseksi, **tunnettu** siitä, että laitteisto sisältää

- muistiyksiköt (101, 102) segmenttejä sisältävän tietämyskannan, hakuindeksien, tietojen ja syötetietovirran käsiteltävän osan tallentamiseksi,

- välineet (102, 103, 106) syötetietovirran lukemiseksi,

- välineet (103, 104, 105) syötetietovirran jakamiseksi elementteihin,

- välineet (103, 104, 105) syötetietovirran ryhmittelemiseksi elementtejä sisältäviin segmentteihin,

- välineet (103, 104, 105) syötetietovirran elementtien analysoimiseksi ja segmenttikohtaisen luokituksen tuottamiseksi analysointitulosten perusteella,

- välineet syötetietovirran segmenttien luokituksen vertaamiseksi tietämyskannan segmenttien luokituksiin ja toisiaan vastaavien segmenttien assosioimiseksi toisiinsa, ja

- välineet (514) segmenttien luokittelun raportoimiseksi.

19. Patenttivaatimuksen 18 mukainen laitteisto, **tunnettu** siitä, että laitteisto sisältää lisäksi välineet (103, 104, 105) syötetietovirran segmenttien vertaamiseksi tietämyskannan segmentteihin.

20. Patenttivaatimuksen 18 mukainen laitteisto, **tunnettu** siitä, että laitteisto sisältää lisäksi välineet (101, 103, 106) vastinelementtejä sisältävien vastinsegmenttien tuottamiseksi jonona, joka muodostaa tulosvirran.

21. Patenttivaatimuksen 18 mukainen laitteisto, **tunnettu** siitä, että laitteistolla on yhteys elementtejä tuottavaan generaattoriin elementtien tuottamiseksi analyysitulosten perusteella.

22. Patenttivaatimuksen 18 mukainen laitteisto, **tunnettu** siitä, että muistiyksiköissä (104, 105) on segmentointitiedot syötetietovirran osan jakamiseksi segmentteihin ja järjestystiedot tulostietovirran segmenttien elementtien järjestyksen määrittämiseksi.

23. Patenttivaatimuksen 18 mukainen laitteisto, **tunnettu** siitä, että muistiyksikössä (104, 105) on tietämyskanta segmenttien, elementtien, luokitusten, vastinsegmenttien ja vastinelementtien tallentamiseksi.

24. Patenttivaatimuksen 18 mukainen laitteisto, **tunnettu** siitä, että laitteistossa on I/O-liityntöjä (106) syöttö- ja tulostietovirtojen lähettämiseksi ja vastaanottamiseksi sekä yhteyden muodostamiseksi muihin järjestelmiin ja/tai käyttäjiin.

25. Patenttivaatimuksen 18 mukainen laitteisto, **tunnettu** siitä, että laitteisto sisältää välineet koko käsiteltävän syötetietovirran osan vertaamiseksi tietämyskannan segmentteihin (606) millä hyvänsä segmenttikoolla.

26. Patenttivaatimuksen 18 mukainen laitteisto, **tunnettu** siitä, että laitteisto sisältää välineet matemaattisten ilmaisujen lukemiseksi ja käsittelemiseksi.

27. Patenttivaatimuksen 18 mukainen laitteisto, **tunnettu** siitä, että laitteisto sisältää välineet formaalien kielien lukemiseksi ja käsittelemiseksi.

28. Patenttivaatimuksen 18 mukainen laitteisto, **tunnettu** siitä, että laitteisto sisältää

- välineet (102, 103, 106) luonnollisen kielen lukemiseksi,

- välineet (103, 104, 105) luonnollisen kielen jakamiseksi elementteihin, jotka ovat sanoja liitteineen,

- välineet (103, 104, 105) luonnollisen kielen ryhmittelemiseksi segmentteihin, jotka ovat sanoja sisältäviä kokonaisuuksia,

- välineet (103, 104, 105) luonnollisen kielen käsiteltävän osion luokittelemiseksi leksikaalisen, morfologisen, syntaktisen tai semanttisen analyysin perusteella, ja

- välineet (101, 103, 106) vastinsanoja sisältävien vastinsegmenttien tuottamiseksi.

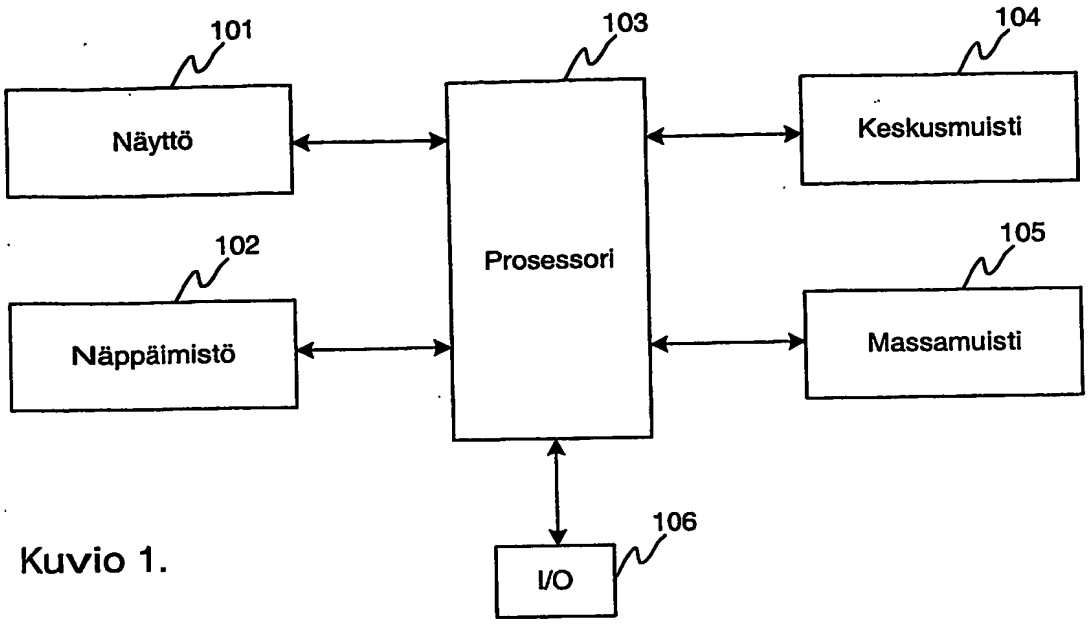
29. Patenttivaatimuksen 28 mukainen laitteisto, **tunnettu** siitä, että laitteistolla on tietoliikenneyhteys vastaavaan laitteistoon jonkin osatoiminnon suorittamiseksi.

**(57) Tiivistelmä**

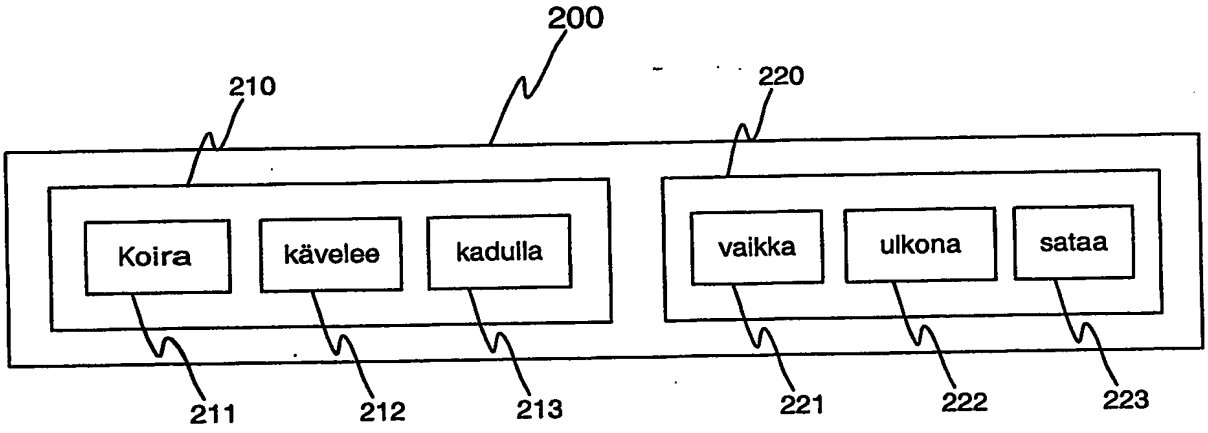
Keksintö koskee menetelmää ja laitteistoa elementtejä (211, 212, 213, 221, 222, 223) sisältävän syötetietovirran (200) datan luokittelemiseksi segmenttejä sisältävän tietämyskannan avulla. Keksinnön soveltuu erityisesti kielien kääntämiseen.

Menetelmässä luetaan (501) käsiteltävä osa syötetietovirrasta (200), jaetaan se elementteihin (211, 212, 213, 221, 222, 223) ja ryhmitellään käsiteltävä osa syötetietovirtaa (200) segmenteiksi (502) siten, että jokainen segmentti (210, 220) sisältää yhden tai useampia elementtejä (211, 212, 213, 221, 222, 223). Käsiteltävän syötetietovirran osan elementit analysoidaan ja analyysitulosten perusteella tuotetaan segmenttikohtainen luokitus. Segmentin luokitusta verrataan tietämyskannan segmenttien (31, 32) luokituksiin ja toisiaan vastaavat segmentit assosioidaan toisiinsa. Tämän jälkeen raportoidaan luokittelun tulos, joka on käsiteltävään syötetietovirtaan assosioitu joukko tietämyskannassa olevia segmenttejä.

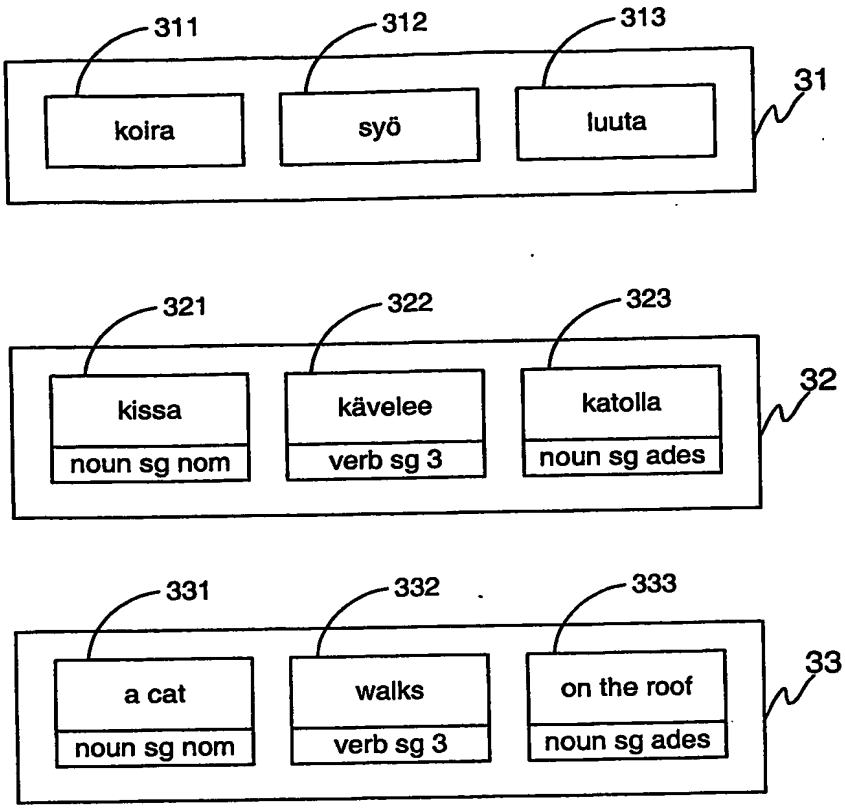
[Kuvio 1]



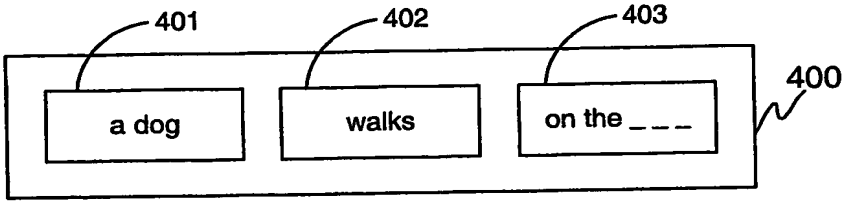

Kuvio 1.



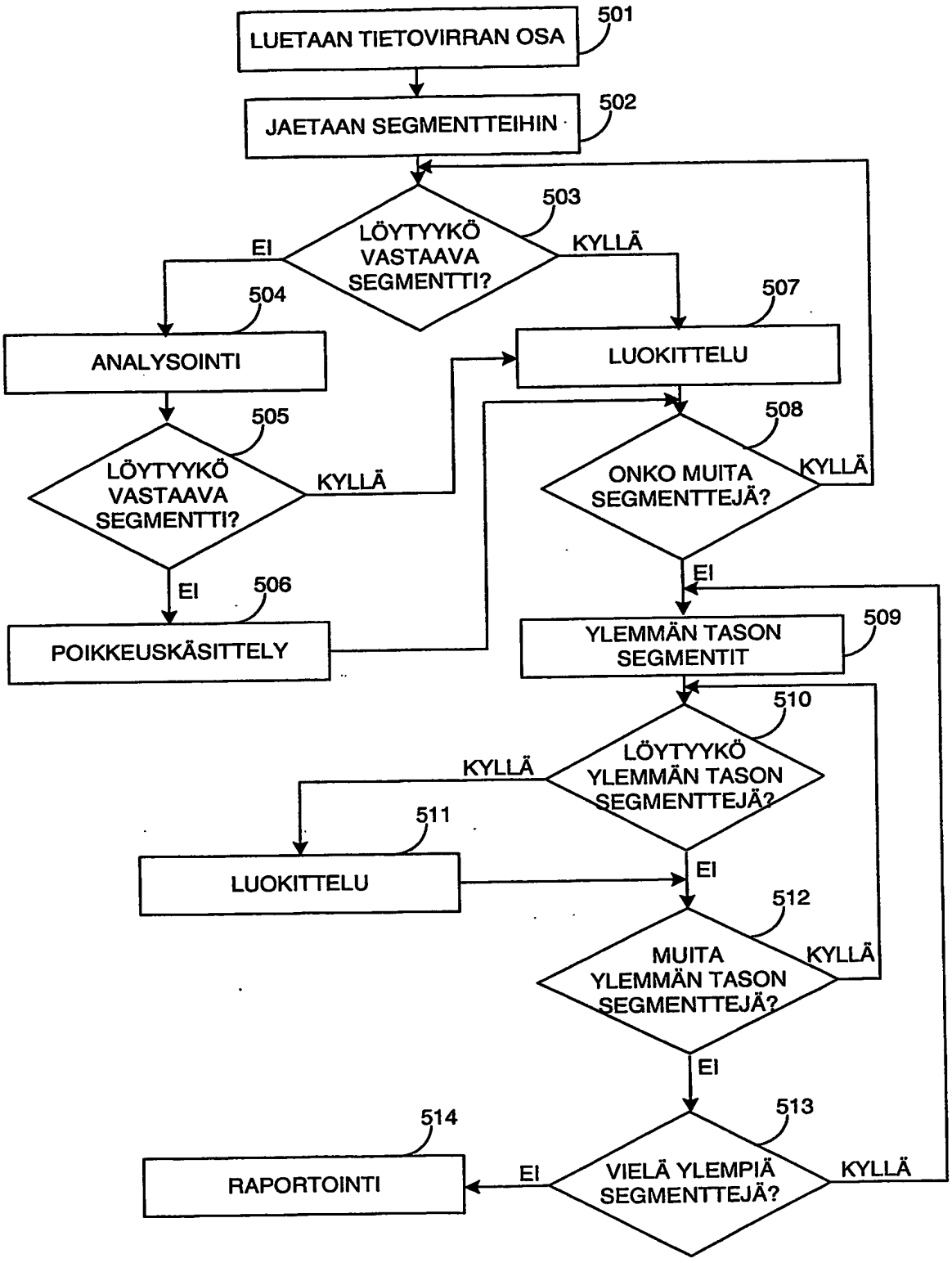

Kuvio 2.



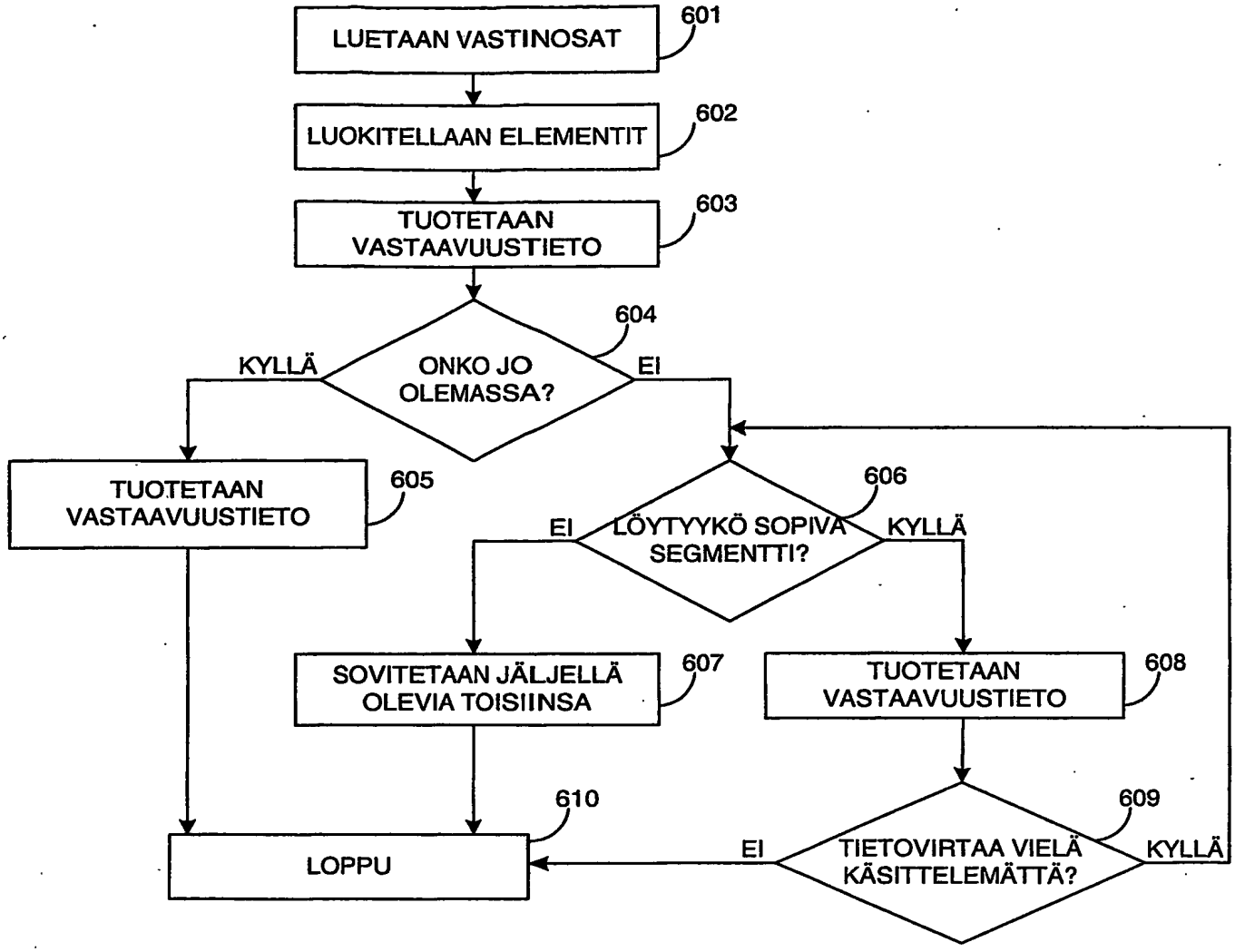

Kuvio 3.



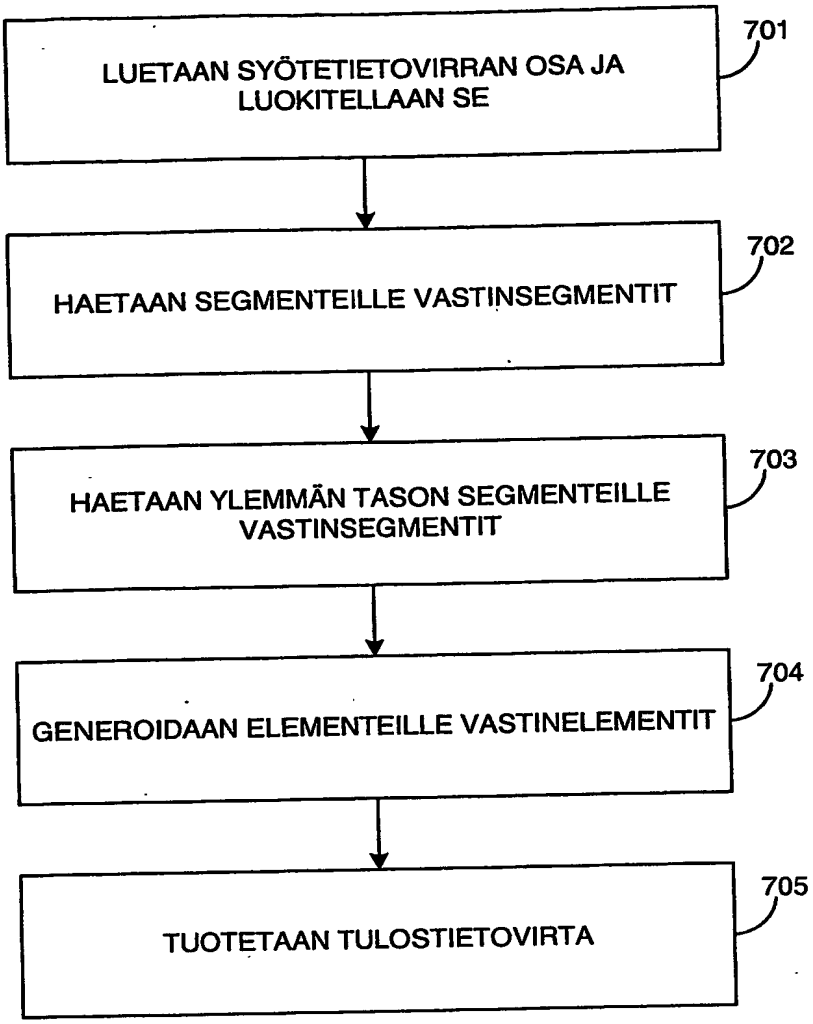

Kuvio 4.


LUETAAN TIETOVIRRAN OSA
501
JAETAAN SEGMENTTEIHIN
502
LÖYTYYKÖ VASTAAVA SEGMENTTI?
503
EI
KYLLÄ
ANALYSOINTI
504
LÖYTYYKÖ VASTAAVA SEGMENTTI?
505
KYLLÄ
EI
POIKKEUSKÄSITTELY
506
LUOKITTELU
507
ONKO MUITA SEGMENTTEJÄ?
508
KYLLÄ
EI
YLEMMÄN TASON SEGMENTIT
509
LÖYTYYKÖ YLEMMÄN TASON SEGMENTTEJÄ?
510
KYLLÄ
LUOKITTELU
511
EI
MUITA YLEMMÄN TASON SEGMENTTEJÄ?
512
KYLLÄ
EI
VIELÄ YLEMPIÄ SEGMENTTEJÄ?
513
KYLLÄ
EI
RAPORTOINTI
514

Kuvio 5.


LUETAAN VASTINOSAT
601
LUOKITELLAAN ELEMENTIT
602
TUOTETAAN VASTAAVUUSTIETO
603
ONKO JO OLEMASSA?
604
KYLLÄ
EI
TUOTETAAN VASTAAVUUSTIETO
605
LÖYTYYKÖ SOPIVA SEGMENTTI?
606
EI
KYLLÄ
SOVITETAAN JÄLJELLÄ OLEVIA TOISIINSA
607
TUOTETAAN VASTAAVUUSTIETO
608
TIETOVIRTAA VIELÄ KÄSITTELEMÄTTÄ?
609
EI
KYLLÄ
LOPPU
610

Kuvio 6.


LUETAAN SYÖTETIETOVIRRAN OSA JA LUOKITELLAAN SE
701
HAETAAN SEGMENTEILLE VASTINSEGMENTIT
702
HAETAAN YLEMMÄN TASON SEGMENTEILLE VASTINSEGMENTIT
703
GENEROIDAAN ELEMENTEILLE VASTINELEMENTIT
704
TUOTETAAN TULOSTIETOVIRTA
705

Kuvio 7.